Efficient prediction of trait judgments from faces using deep neural networks

Umit Keles[1] ✉, **Chujun Lin**[2], **and Ralph Adolphs**[1,3]

[1]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA
[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA

Judgments of people from their faces are often invalid but influence many social decisions (e.g., legal sentencing), making them an important target for automated prediction. Direct training of deep convolutional neural networks (DCNNs) is difficult because of sparse human ratings, but features obtained from DCNNs pre-trained on other classifications (e.g., object recognition) can predict trait judgments within a given face database. However, it remains unknown if this latter approach generalizes across faces, raters, or traits. Here we directly compare three distinct types of face features, and test them across multiple out-of-sample datasets and traits. DCNNs pre-trained on face identification provided features that generalized best, and models trained to predict a given trait also predicted several other traits. We demonstrate the flexibility, generalizability, and efficiency of using DCNN features to predict human trait judgments from faces, providing an easily scalable framework for automated prediction of human judgment.

social cognition | face perception | deep learning | neural networks

Correspondence: ***ukeles@caltech.edu***

People rapidly and spontaneously make trait judgments about unfamiliar others based on their faces, such as forming an impression that someone looks beautiful, trustworthy, or threatening [1–3]. These judgments have ubiquitous and major consequences in everyday life. For instance, a large body of research has demonstrated that trait judgments of political candidates based merely on faces (e.g., how competent an unfamiliar candidate looks) are associated with election outcomes across various regions of the world [4–6], with evidence suggesting that these trait judgments causally influence individual voting decisions [7,8]. Other examples of trait judgments influencing real-life decisions range from picking out dates, to hiring employees, choosing science news, and determining courtroom sentences [9–12]. In the real world, these judgments can show large individual differences and context effects: not only are they invalid, but consensus can be difficult to achieve even for stimuli argued to be universal, such as emotional facial expressions [13]. With these constraints in mind, it remains a fact that people nowadays do make many judgments solely from faces in the absence of context or other information (e.g., deciding not to date someone just based on profile photos on dating sites), and such judgments show considerable consensus across cultures [14].

An important applied question is whether machines could be trained to make trait judgments from faces like humans do. Recent work has trained deep convolutional neural networks (DCNNs) on face images that had been previously rated on various traits to predict how humans would judge new face images on the same set of traits [15,16]. While this approach is informative, it turns out to be unnecessary. DCNNs that have only been trained to recognize face identity, or even object identity, without any training specifically on trait judgments, already generate features that can be used in simple regression models to predict human trait judgments of faces [17,18]. This finding is perhaps unsurprising, since, in the absence of any other context, the structural features of the face are also the only source of information that human raters have available for their trait judgments. This approach in principle offers a more flexible and scalable framework for practical application: new faces can be projected into the same, pre-trained DCNN to generate facial features, which could then be used in regression models to predict trait judgments. This takes advantage of the power of existing pre-trained DCNNs that typically generalize over pose, viewpoint, and image quality, and obviates the need to train new DCNNs or retrain existing networks on domain-specific trait ratings, which is inefficient [19,20].

Past and current work highlights several specific limitations of using DCNNs to predict human trait ratings. First, inconsistent results have been found when using features from DCNNs pre-trained for face identification versus those for object recognition [17,18]; it is also unclear how features from different pre-trained DCNNs explain the variance in trait judgments of faces. Second, all prior studies trained and tested their models using a single dataset (the 10k US Adult Face Database [21] in Song et al. [17], and ratings for the Human ID Database [22] in Parde et al. [18]), leaving it an open question how well this approach generalizes out-of-sample (both across face databases and across human raters), a growing concern in modern machine learning for practical applications [23]. Third, recent findings show that human judgments of faces on a large number of traits can be captured by a small number (two to four) of psychological dimensions [14,24,25], raising the possibility of generalizability across traits (a model trained to predict a particular trait from face features should also predict judgments of some other traits), an important shortcut that remains to be tested empirically.

We address the above three open questions in the present study to provide a robust, generalizable, and efficiently scalable framework for automating trait inferences from faces. We trained regularized linear regression models with cross-validation to predict human trait judgments of faces using features from three distinct spaces (Fig. 1a-b; see also Methods): a pre-trained DCNN for face identification [26]

(*DCNN-Identity*), a pre-trained DCNN for object recognition[27] (*DCNN-Object*), and facial landmarks (*Landmark*; e.g., eye size[28]; Supplementary Fig.1) for comparison to previous findings (e.g., faces with wider eyes are perceived as more honest[29]). We elucidate how the three distinct feature spaces explain the variance in trait judgments of faces. All models were trained primarily on the neutral, frontal, white faces (n = 183) and their corresponding available trait ratings from the Chicago Face Database[28], a widely used database in machine learning studies of faces. To characterize the generalizability of the current approach across faces, raters, and traits, we tested the models in six out-of-sample datasets that included ratings for different types of face images on a variety of traits provided by independent samples of human subjects (Fig. 1c).

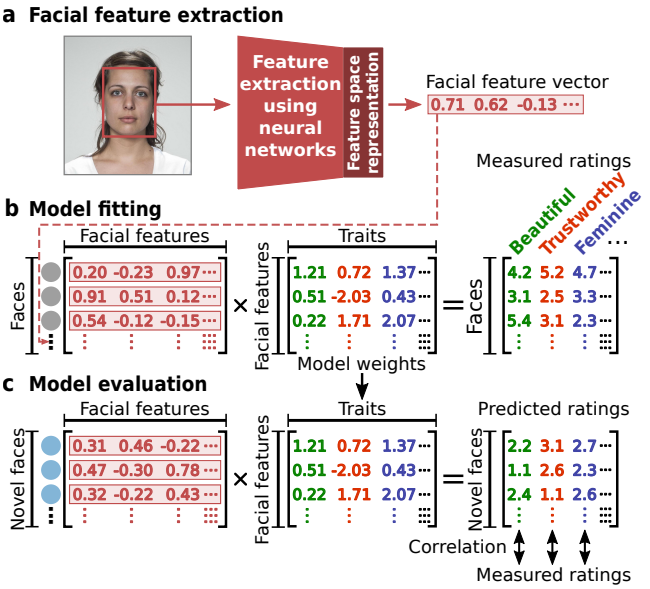


Figure 1. Overview of modeling framework. **a,** Face images were projected onto three distinct feature spaces: a feature space obtained from the top layer of a pre-trained DCNN for face identification (DCNN-Identity)[26]; a feature space obtained from the block5_conv2 layer[17] of a pre-trained DCNN for object recognition (DCNN-Object)[27], and a feature space obtained from geometric measures of the faces (Landmark)[28]. **b,** Regularized linear regression with cross-validation was used to estimate a set of model weights for each trait, which maps each feature space onto the trait ratings. **c,** The estimated model weights were then used to predict the trait ratings for novel faces from their facial features. Models constructed for the three distinct feature spaces were compared based on how accurately they predicted ratings for novel faces (Spearman's correlation between the predicted trait ratings and the actual trait ratings collected from human participants).

Results

Generalizability across faces and raters. For each trait and each facial feature space, we trained a regularized linear regression model with cross-validation to learn the relationship between the features and human judgment of this trait from faces. Results reported here were from models trained on the popular Chicago Face Dababase[28]; we also trained the models using a more recent database that collected ratings from human subjects on a much larger number of traits for representatively sampled faces[14], whose results corroborated those reported here (see Supplementary Fig. 2; the trait models we trained for automated prediction were not identical across the two training datasets because the two datasets differed in the traits for which human subject ratings were available).

To investigate how well the predictions of our modeling approach generalized across faces and raters, we tested the models on multiple out-of-sample datasets that collected trait ratings from independent sets of human subjects for different types of novel faces (studio portraits, ambient photos that varied in viewpoint, facial expression, background, etc.) than the training set (studio portraits only). Since the traits with available human ratings in the test datasets were not identical to those in the training dataset, we only computed the prediction accuracy for those traits in the test datasets that were the same or semantically highly similar (or the exactly opposite) to those in the training dataset (e.g., predicting *submissive* ratings in the test dataset using the models trained on the *dominant* ratings in the training dataset by multiplying the model weights with -1). Results summarized in Fig. 2 showed that the *DCNN-Identity* models predicted almost all traits across all datasets (except *dominant* ratings for ambient photos, Fig. 2d), and yielded a higher prediction accuracy across traits and test datasets (Spearman's correlations, M = 0.55, SD = 0.19 across traits and test datasets) than the *DCNN-Object* models (M = 0.43, SD = 0.22) or the *Landmark* models (M = 0.38, SD = 0.21).

The superior performance of the *DCNN-Identity* models over the *DCNN-Object* and *Landmark* models raise two questions: whether it was simply due to the much larger number of features in the *DCNN-Identity* models (n = 128), and whether it was idiosyncratic to the specific network used to derive those *DCNN-Identity* features. To address the first question, we applied principal component analysis (PCA) on the *DCNN-Identity* features and used only the first 30 PCs for fitting the models, a number close to the number of features in the *DCNN-Object* (n = 26) and the *Landmark* models (n =30). To address the second question, we fit the models using features from a different DCNN for face identification that has a distinct architecture than the *DCNN-Identity* network, the OpenFace DCNN[32]. Results showed that the performance of the *DCNN-Identity* PC models was as good as the original *DCNN-Identity* models, and the superior performance of the original *DCNN-Identity* models was not idiosyncratic to the specific network architecture (Supplementary Fig. 3).

Comparison across feature spaces. We have shown that models using *DCNN-Identity* features predicted trait ratings from faces at a higher accuracy than models using the other two feature spaces across various traits and test datasets. We next sought to quantify the variance explained by models using each of these three feature spaces. We performed a variance partitioning analysis[33,34] (see Methods) to identify the proportion of unique variance in the trait ratings that was explained by each feature space and the proportion of shared variance that was explained by multiple feature spaces. Models were trained and tested using the training dataset and test dataset as in Fig. 2a.

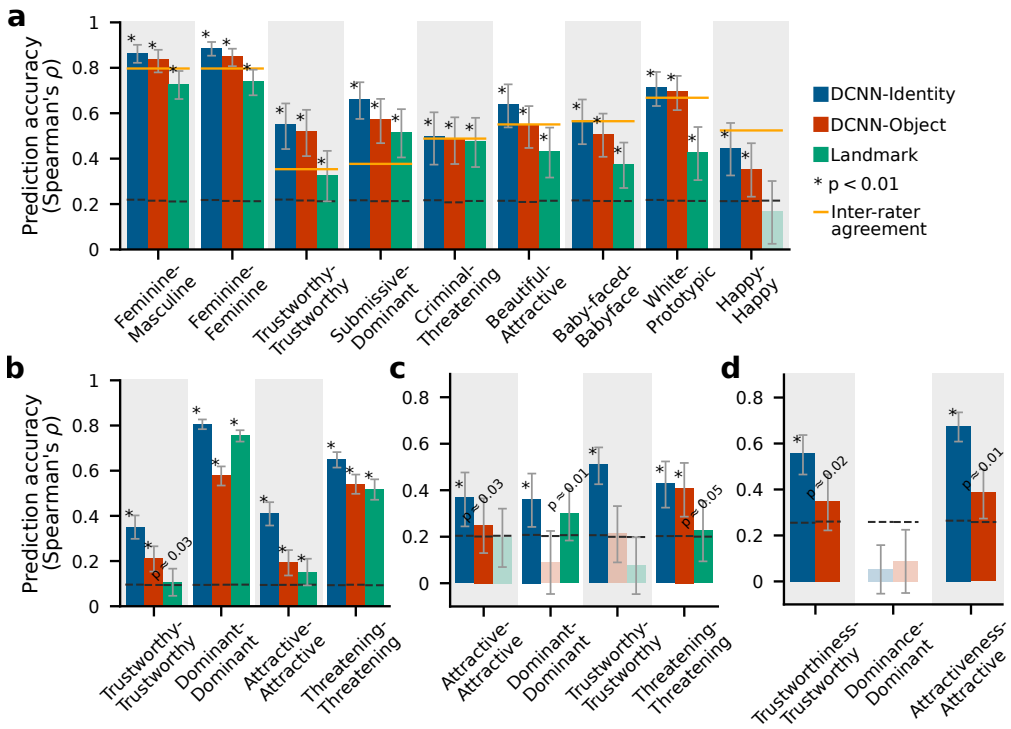


Figure 2. Prediction accuracy of three feature spaces across different test datasets. All models were trained on human subject ratings for 183 studio portraits of frontal, neutral, white faces from the Chicago Face Database[28]. The x-axis indicates the trait(s) measured in the test and training datasets (test-training). **a,** The prediction accuracy of the models tested on an independent dataset of 60 novel studio portraits of frontal, neutral, white faces and their trait ratings[14]. The bar height indicates the mean prediction and error bars indicate the standard deviations of the mean prediction accuracy across bootstrap samples. Orange lines indicate the average inter-rater agreements in the test data (the average of the Spearman correlations between ratings from all pairs of individual participants). **b,** The prediction accuracy of the models tested on a different independent dataset of 300 computer-generated white faces and their trait ratings[24]. **c,** The prediction accuracy of the models tested on a different independent dataset of 66 studio portraits of frontal, neutral, white faces and their trait ratings[30]. **d,** The prediction accuracy of the models tested on a different independent dataset of 504 ambient photos of white faces in the wild (varied in viewpoint, facial expression, background, illumination, etc.) and their trait ratings[31] (42 images were used in each bootstrap iteration, see Methods). The automatic extraction of Landmark features was not feasible for these faces. Saturated colors, asterisks, and p-values indicate statistically significant predictions ($p < 0.05$, assessed with permutation tests, and FDR corrected); desaturated colors indicate insignificant predictions. Dashed black lines indicate the chance threshold ($p = 0.05$, assessed with permutation test) for the prediction accuracy.

The variance partitioning analysis revealed that the *DCNN-Identity* and *DCNN-Object* models accounted for almost the same variance in test datasets (Fig. 3a). The *Landmark* model, on the other hand, was not able to explain any unique variance beyond that shared with the other two feature spaces (Fig. 3b-c). Thus, all the variance in the trait ratings that was explained by any of the three feature spaces could be explained by the *DCNN-Identity* features alone.

The highly similar explained variance between the *DCNN-Identity* and the *DCNN-Object* feature spaces raises an interesting question: given the same face image, are the facial features that are relevant for predicting human trait judgments extracted using the *DCNN-Identity* network highly similar to those extracted using the *DCNN-Object* network? To provide insights into this question, we manipulated the face images in the test dataset on a list of low-level image properties–their color, hair region, and mean luminance (Supplementary Fig. 4a-d)–changes that we expect to have minimal impact on human trait judgments. We used the previously trained regression model weights (i.e., models trained on the unmanipulated version of the faces as in Fig. 2a) and the features of the manipulated versions of the face images in the test dataset (extracted using the *DCNN-Identity* network and the *DCNN-Object network*, respectively) to predict the human trait ratings of the unmanipulated version of the face images.

We found that the manipulation of these low-level image properties yielded a larger decline in the prediction accuracy of the *DCNN-Object* models ($M = 0.28$ across traits, especially in the predictions of *trustworthy*, *criminal*, *white*, and *happy*), but only a slight drop in the prediction accuracy of the *DCNN-Identity* model ($M = 0.05$ across traits) as shown in Supplementary Fig. 4e-f. These results suggest that the *DCNN-Object* features carry substantial information about image-based characteristics (e.g., illumination, hair parts close to face area), limiting the generalizability of predicting human trait judgments of the same face in different image styles; whereas, the *DCNN-Identity* features carry face-specific information (e.g., identity, gender) that is largely robust to the changes in image styles[19,20]. Taken together, these results indicate that DCNNs pre-trained to recognize face identity produce features that can be used most successfully to predict trait judgments made by humans from faces, and that these predictions generalize well across faces, raters, and image styles.

Generalizability across traits. Recent research has shown that the hundreds of different trait words people use to describe judgments of others from faces could be represented by just a few trait dimensions[14,24,25] (typically 2-4 dimensions account for > 70% of the variance in ratings). These

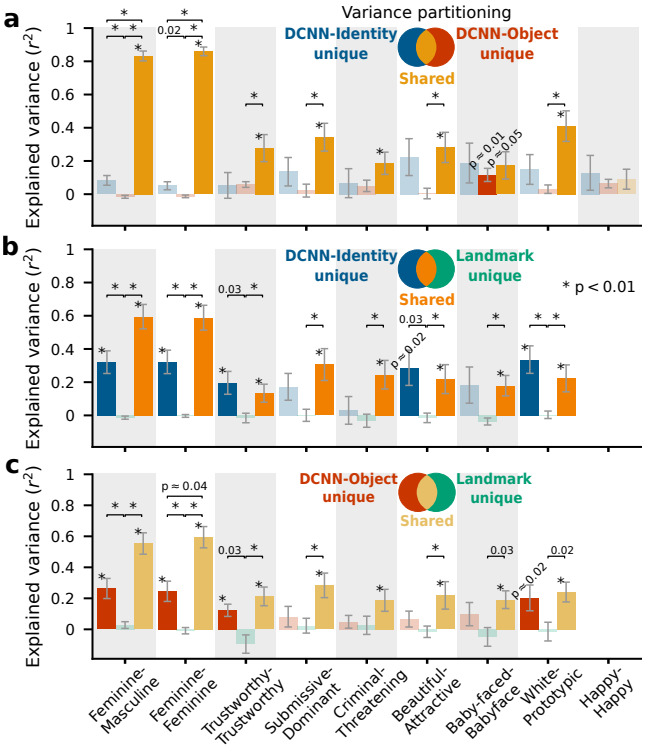


Figure 3. Results of variance partitioning analyses. a, Variance partitioning between the DCNN-Identity and DCNN-Object models. Error bars show bootstrap standard deviations of the explained variance. Saturated colors, and the asterisks and p-values next to the error bars indicate that the explained variance was significantly different from zero ($p < 0.05$, assessed with bootstrap tests, FDR corrected). Desaturated colors indicate that the explained variance was not significantly different from zero. The asterisks and p-values above the horizontal brackets indicate significant differences in the explained variance between unique and shared components ($p < 0.05$, bootstrap tests, FDR corrected). **b,** Variance partitioning between the DCNN-Identity and Landmark models (this analysis was not performed for the trait happy because the Landmark model failed to predict this judgment in the test dataset, see Fig. 2a). **c,** Variance partitioning between the DCNN-Object and Landmark models.

findings highlight the possibility that the individual models trained for different traits would also be correlated[35,36]. Indeed, we found that traits that were correlated in the original human subject ratings across face images were also correlated in their model weights across features (Fig. 4). None of the correlations computed with the human subject ratings was significantly different from the correlation computed with the estimated model weights for the same pair of traits (bootstrap tests, $p > 0.05$, FDR corrected).

These results suggest that it is not necessary to train an individual model for every different trait, and that it is possible to obtain model predictions even for traits whose human ratings were unavailable in the training dataset. Therefore, we next investigate to which degree a model trained to predict a certain trait judgment from faces (e.g., *feminine*) would generalize to predict the judgments of other traits (e.g., *sociable*) regarding the same face ("cross-prediction"). We assessed the cross-prediction accuracy with Spearman correlation between the ratings predicted by the model for a certain trait (e.g., *feminine*) and the ratings collected from human subjects for a different trait (e.g., *sociable*) regarding the same set of faces in the test dataset. All analyses in this section used the same training dataset and test datasets as in Fig. 2 (with one additional test dataset[37]). We found a large number of accurate cross-predictions in all test datasets (Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e). For instance, the model trained to predict *feminine* judgments from faces not only predicted *feminine* judgments in the test dataset as intended (Fig. 5a and Fig. 2a), but also predicted how much human subjects judged the faces in the test dataset to be *beautiful*, *sociable*, *submissive*, *trustworthy*, etc. (Fig. 5a).

Given that most trait judgments from faces were cross-predicted by multiple trait models (e.g., per row in Fig. 5a) and each trait model cross-predicted different trait judgments (e.g., per column in Fig. 5a), we next investigate which trait model plays the most important role in cross-prediction–thus, most generalizable for automatic prediction across traits. For each cross-prediction (e.g., using the *feminine* model to predict ratings of *sociable* in the test dataset), we computed the residual cross-prediction accuracy after partialling out the effect of each remaining trait model (i.e., the eight trait models in the x-axis of Fig. 5a except for *feminine*). The residual cross-prediction accuracy (Fig. 5b, Supplementary Figs. 5b, 6b, 7b,d,f) was assessed with the semi-partial Spearman's correlation between the human ratings of a trait for the faces in a test dataset (e.g., *sociable*) and the residuals from a simple bivariate regression of the predicted ratings of a trait model for faces in the same test dataset (e.g., *feminine* model) on the predicted ratings of a remaining trait model for the same faces (e.g., *trustworthy* model; these residuals quantify the unique variance in the predicted *feminine* ratings that were not associated with the predicted *trustworthy* ratings). Figure 6 summarizes the mean residual cross-prediction accuracy in each test dataset after partialling out the effect of each remaining trait model. We found that models predicting gender (*masculine/feminine*) played a more important role for cross-prediction of personality traits from faces (Big-2 and Big-5) and trait judgments of computer-generated faces[24,37]. By contrast, models predicting *trustworthy* and *happy* played a more important role in test datasets where the photos were neutral and taken for research purposes[14,30], and model predicting *attractiveness* was more important for ambient social media profile photos[31].

Discussion

In this paper, we trained regularized linear regression models (Ridge regression with cross-validation) to make trait judgments from faces based on the features extracted from the faces using pre-trained DCNNs (Fig. 1). We investigated the generalizability of this approach to out-of-sample faces and raters (Fig. 2) and across traits (Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e) using six independent test datasets.

We found that regression models using features from DCNNs that were pre-trained to distinguish facial identity (*DCNN-Identity*) predicted human trait judgments from faces most accurately and generalized the best across faces and raters (Fig. 2) compared to the models using DCNN features for object recognition (*DCNN-Object*) or features based on facial landmarks (*Landmark*). The superior performance of

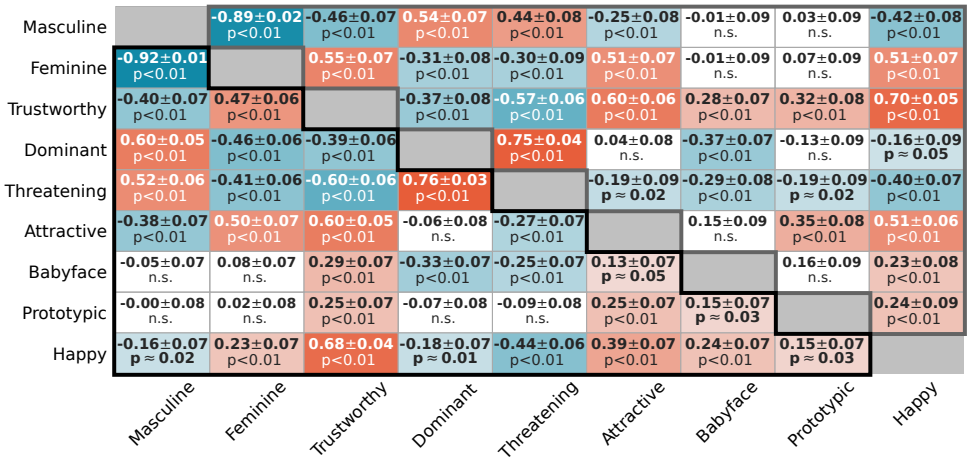


Figure 4. Correlations between traits in human subject ratings and estimated model weights. The lower-triangle panel shows the Spearman correlations among traits computed using the human subject ratings across face images per trait in the training dataset. The upper-triangle panel shows the Spearman correlations among traits computed using the estimated model weights across features per trait in the training dataset. The saturation of color indicates the magnitude of the correlation (red for positive, blue for negative). Numbers indicate the mean, standard deviation, and significance of the correlation (bootstrap tests, FDR corrected).

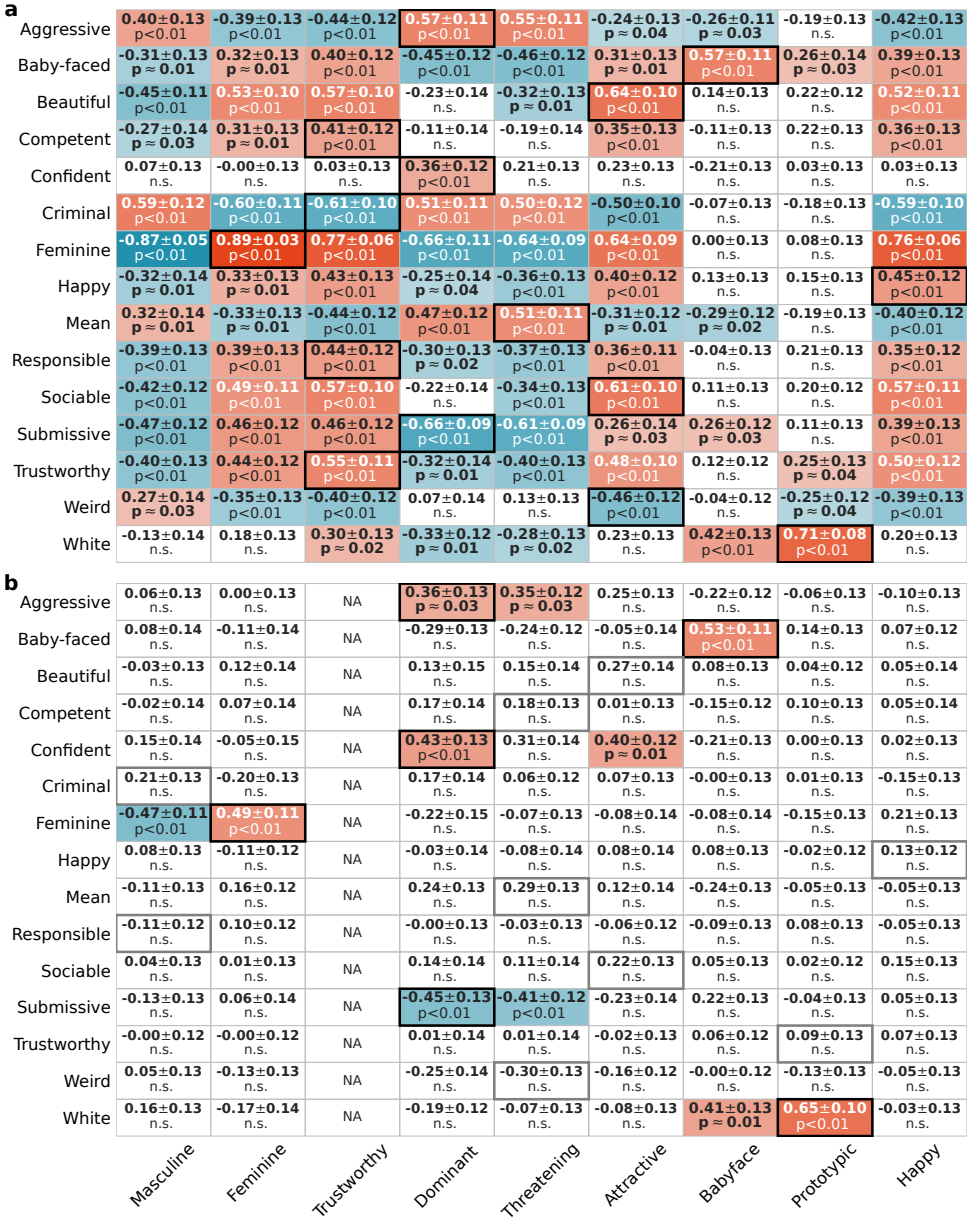


Figure 5. Cross-prediction accuracy across traits. a, Cross-prediction accuracy (the Spearman correlations) between the predicted ratings of the faces in the test dataset used in Fig. 2a on nine traits (x-axis) and the human subject ratings of the same set of faces on fifteen traits (y-axis). The saturation of the color indicates the magnitude of the correlation (red for positive, blue for negative). Numbers indicate the mean and standard deviation (across bootstrap samples), and the significance of the correlation (permutation test, FDR corrected). **b,** An example of residual cross-prediction accuracy for traits in the test dataset used in Fig. 2a (y-axis) from eight trait models (x-axis) while controlling for the prediction from the trustworthy model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). Numbers report the mean bootstrap residual cross-prediction accuracy, bootstrap standard deviation, and significance level computed via permutation tests and FDR corrected. The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes); the highest accuracy per row was highlighted with a solid box (black for significant, grey for insignificant).

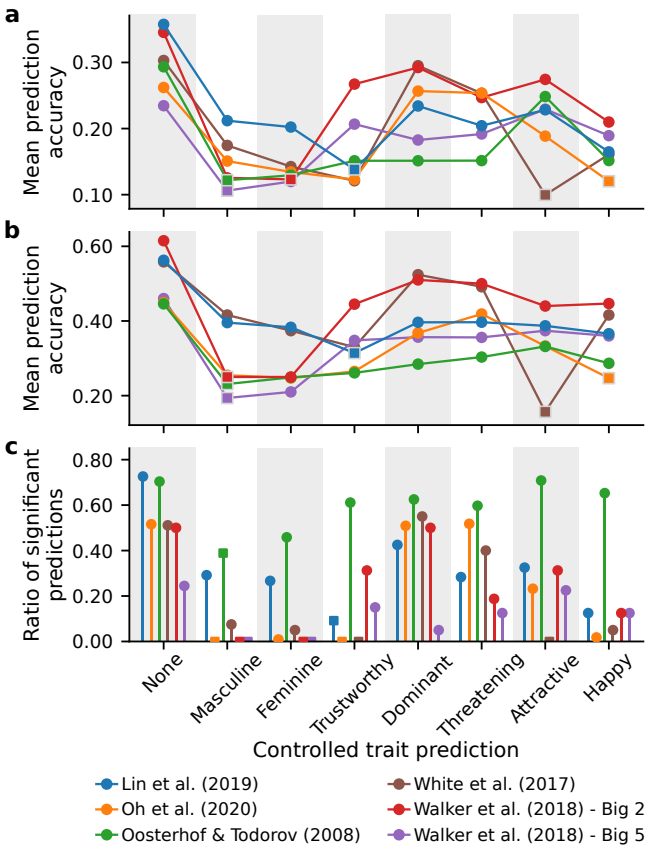


Figure 6. Residual cross-prediction accuracy after partialling out the effect of a second trait model. **a,** The first column ("None") plots the mean cross-prediction accuracy (dots; i.e., mean absolute Spearman correlations) across all cross-predictions in each test dataset (all cells in Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e). The other columns plot the mean residual cross-prediction accuracy across all cross-predictions after partialling out the effect of the trait model labeled in the x-axis. The square indicates the trait model (x-axis) that was the most impactful for cross-predictions in a test dataset (i.e., minimum mean residual cross-prediction accuracy). **b,** The first column ("None") plots the average maximum cross-prediction accuracy (the maximum per column in Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e) across all trait models (x-axis in Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e). The other columns plot the average maximum residual cross-prediction accuracy across all trait models after partialling out the effect of the trait model labeled in the x-axis. **c,** The first column ("None") plots the ratio of significant cross-predictions across all cross-predictions in each test dataset. The other columns plot the ratio of significant cross-predictions after partialling out the effect of the trait model labeled in the x-axis.

the *DCNN-Identity* models were robust to the dataset used to train the models (Supplementary Fig. 2), the number of features included in the regression models, and the network architecture used to obtain the identity features (Supplementary Fig. 3).

Using variance partitioning analysis, we showed that the *DCNN-Identity* models and the *DCNN-Object* models explain almost the same variance in the trait ratings (Fig. 3). However, the features extracted by the two networks from faces that were relevant for predicting human trait judgments of the faces were not the same, with the former representing more information unique to faces (e.g., identity, gender) rather than images in general and being more robust to manipulations of image styles (Supplementary Fig. 4).

Our cross-prediction analysis showed that it was not necessary to train an individual model for each different trait judgment, which is a common practice in computational modeling of trait judgments, but that prediction was feasible even for traits whose human subject ratings were not available in the training dataset for training a prediction model (Fig. 5a, Supplementary Figs. 5a, 6a, 7a,c,e). These findings are in line with recent psychology research on the low-dimensional representation of trait judgments from faces [14,24,25]. We also provided a novel analysis, semi-partial correlation analysis, for understanding the important trait dimensions for automatic predictions (e.g., Fig. 5a). This analysis could potentially be applied to a broader set of trait models to help understand the important trait dimensions humans use to make trait judgments from faces in different contexts (e.g., photos taken for different purposes, or for different types of decision-making). For instance, we found that the judgments of *trustworthy* and *happy* seem to play a more important role in test datasets with neutral faces for research purpose, whereas the judgment of *attractiveness* was more important in the test dataset where the photos were taken for social media profiles by the user themselves (Fig. 6).

Taken together, these results showed that the approach of training simple linear regression models using features from pre-trained facial identity DCNNs offers a flexible framework for predicting human trait judgments from faces. This approach works well even for ambient photos of faces in the wild, for which the traditional facial metrics (*Landmark*) are difficult to measure, and for predicting judgments of correlated traits whose human subject ratings might not be available in the training data.

The most important limitation of our findings is that they almost certainly lack validity. That is, even though there is considerable consensus in the trait judgments made by humans from faces [38] (generating the "ground-truth" labels for training trait models), the majority of those judgments do not reflect the actual traits of the person whose face is shown. Instead, those judgments mainly reveal our biases and stereotypes [39,40]. This limitation is even more acute given that all stimuli in both the training and test datasets were isolated faces devoid of context and any other information about the person. Our results thus show that it is possible to predict what people judge or believe about brief glances of a face, but not what is in fact valid about the person whose face is shown as the stimulus. Needless to say, it is critical to keep this distinction in mind: we did not predict anything about the people whose faces were used; instead, we predicted what human viewers judge about those faces.

Our conclusions were also limited by the small number of overlapping traits between the training dataset and the different test datasets, and thus the small number of traits we could build models for in the present paper. Finally, we only included white faces in our analysis, since these were the predominant race available in the training datasets. This restricts our findings, since previous work has shown that human trait judgments are influenced by the unique facial features of faces from different races (via a bottom-up psychological process) as well as the different social concepts associated with the different races (via a top-down psychologi-

cal process)[41,42]. Altogether, the restricted range of different types of faces, and the small number of trait ratings, provide results that are not yet comprehensive. Analyses on future datasets that are more complete will be valuable to extend the present study.

We conclude with two future directions. First, our findings show that future research and application could flexibly train a set of models using a particular dataset (as we used the Chicago Face Database here) and predict trait ratings collected in entirely different datasets for different types of faces and from different groups of human subjects (as we used six different independent test datasets here), or train a set of models for certain traits and predict ratings of different sets of correlated traits. Second, trait judgments are commonly thought to reflect temporally stable attributes linked to the identity of a person; whereas emotion judgments are relatively independent of identity, and instead are based on transient changes in facial muscles. Nonetheless, emotional expressions are found to influence certain trait judgments[43,44]. Future research combining features from DCNNs pre-trained for face identification and for emotion categorization could further improve prediction accuracy for human trait judgments of faces.

Methods

Training and test datasets. The data used in the present research were from publicly available datasets and previously published studies. The models were trained on 183 studio portraits of neutral, frontal, white faces and their trait ratings from the Chicago Face Database[28]. Faces in this database that are not neutral, not frontal, and of other races were excluded since the effects of those factors are beyond the scope of our current research. Each face image was rated by human subjects on fifteen traits (*afraid*, *angry*, *attractive*, *baby-faced*, *disgusted*, *dominant*, *feminine*, *happy*, *masculine*, *prototypic*, *sad*, *surprised*, *threatening*, *trustworthy*, and *unusual*) using a 1-7 Likert scale (1 = Not at all, 7 = Extremely). In the current study, we focused on judgments related to temporally stable attributes, such as physical appearance (e.g., *baby-faced*, *feminine*) and social traits (e.g., *dominant*, *trustworthy*, *threatening*), and excluded most emotional expressions (*afraid*, *angry*, *disgusted*, *sad*, *surprised*) except for *happy* which was also measured in one of the out-of-sample test datasets. We also excluded judgment of *unusual* because it was not measured in any of the test datasets we used.

The models were tested on six independent publicly available datasets[14,24,30,31,37]. These test datasets were selected to sample trait ratings for different types of faces, including studio portraits of frontal, neutral faces, computer-generated faces, and ambient photos of faces taken under unconstrained conditions. All faces in our training and test datasets were limited to white faces; the effects of race and context are beyond the scope of our current study. All ratings used for model training and testing were averaged across human subjects per face per trait.

DCNN-Identity features. To extract identity features from face images, we used the dlib library, which offers an open source implementation of face recognition with deep neural networks[26]. The network represents each face image with a vector of 128 features. The network had been originally trained to identify 7,485 face identities in a dataset of about three million faces with a loss function such that the two face images of the same identity were mapped closer to each other in the face space than the face images of two different identities. Built on a ResNet architecture with 29 convolutional layers, the network achieved a state-of-the-art accuracy of 99.38% on the Labeled Faces in the Wild benchmark[26]. The features extracted from networks trained for face recognition capture facial features related to individual identity, such as face geometry, but ignore factors that could vary for each individual, such as body pose, facial expression, and image luminance[20]. We directly used the feature vectors coming from the last layer of the network, without tuning the network or its last layer specifically for trait judgments from faces.

DCNN-Object features. To extract object features from face images, we used the features obtained from block5_conv2 layer of the VGG16 network based on their success in predicting human trait judgments from faces in a prior study[17]. To extract the features from a face image, the face region of the image was first detected and segmented automatically. Then the segmented image was presented to the VGG16 model implemented in the Keras deep learning library[45] with weights pre-trained on the ImageNet dataset[46] for object recognition. The output of the block5_conv2 layer had a volume shape of $14 \times 14 \times 512$ which was flattened into a 100352-dimensional feature vector. Thus, the layer represented each face image with a vector of 100352 features.

Due to the large number of features, we used principal component analysis (PCA) to reduce the dimensionality and redundancy of these features. To maximize the principal components' relevance for face representation, PCA was performed on a large face database of 426 white adults with neutral expression aggregated from three popular publicly available face databases[28,47,48]. The optimal number of components was determined based on their performance for predicting trait ratings in the model training dataset (i.e., the 183 studio portraits from the Chicago Face Database). Specifically, the 426 faces were first represented with the 100352-dimensional feature vectors, with which we performed PCA to extract PCs of the features; next, the 100352-dimensional feature vectors of the 183 faces in the training dataset were projected onto the PCs obtained from the 426 faces; then, we fit ridge regression models with a nested cross-validation procedure (see Methods: Model fitting) using different numbers of PCs to predict the trait ratings of the faces; the models were trained on a subset of the 183 faces and their ratings, and tested on a held-out subset of data; finally, we examined the average prediction accuracy across traits as a function of the number of PCs (increased from 10 to 80 with a step size of one), which achieved the highest accuracy with 26 PCs.

Landmark features. The physical and geometric features of the face have been shown to influence how humans form trait impressions of unfamiliar others based on faces[28]. To obtain these features, we referred to the 40 facial metrics provided in the Chicago Face Database[28], which were defined based on a review of the social perception literature[49,50] and manually measured using an image editing software[28]. In our present study, given the large number of faces we used, we aimed to generate a subset of those facial metrics that could be automatically measured. Specifically, we used a pre-trained model of facial landmark detection that estimates the location of 68 key points on each face image and a pre-trained model of face parsing that segments each face image into several facial parts such as skin area, left and right eye, nose, etc.[26,51] (see Supplementary Fig. 1). These automated methods allowed us to obtain 30 physical and geometric features (*Landmark* features) that closely imitate the manually measured facial metrics provided in the Chicago Face Database. The 30 *Landmark* features were the median luminance of skin area, nose width, nose length, lip thickness, face

length, eye height (left, right), eye width (left, right), face width at cheek, face width at mouth, distance between pupils, distance between pupil and upper lip (left, right, asymmetry), chin length, length of cheek to chin (left, right), face shape, (face) heartshapeness, nose shape, lip fullness, eye shape, eye size, midface length, chin size, cheekbone height, cheekbone prominence, face roundness, and facial width-to-height ratio. We verified that the 30 automatically extracted *Landmark* features described the trait judgments from faces equally well as the manually measured features by comparing the prediction accuracy of the trait models based on each of the two feature sets respectively (see Supplementary Fig. 1).

Model fitting. L2-regularized linear regression (a.k.a., Ridge)[52] was used to train a set of model weights separately for each trait that optimally mapped facial features onto human trait ratings of the faces (Fig. 1). Cross-validation was used to determine the optimal regularization parameter for Ridge regression. Specifically, the training dataset was randomly split into 80% training and 20% of validation samples for 2000 iterations. At each iteration, a range of regularization parameters (n = 30, log-spaced between 1 and 100,000) were used to fit models to the training part, and each fit model was used to predict the human ratings of the faces in the validation part. This procedure yielded a model accuracy per regularization parameter per iteration per trait, assessed with the coefficient of determination (R^2)[34,53]. For each trait, the optimal regularization parameter that maximized the average accuracy across all iterations was selected, and the model weights were refit using this optimal regularization parameter using the entire training dataset (i.e., the final trait model).

The final trait model was used to predict ratings of the same trait or semantically highly (dis)similar traits for the novel faces in each independent test dataset. A bootstrap procedure was used to robustly estimate the prediction accuracy of each trait model on each test dataset. Specifically, the face images and their trait ratings in each independent test dataset were randomly sampled 10,000 times with replacement, and the Spearman rank-order correlation between the resampled predicted and resampled human trait ratings was computed per trait[33,34]. We used the Spearman rank-order correlation to assess model accuracy because the ratings in some test datasets were collected on a different scale than the training dataset. The mean prediction accuracy for each trait was obtained by averaging the accuracies across bootstrap iterations. For the test dataset that contained a large number of ambient photos (504 photos of 42 white individuals were selected for testing from the 1224 photos of 102 individuals of all races)[31], one image was randomly sampled from each individual's images at each bootstrap iteration (i.e., 42 images were included at each iteration) to prevent bias in prediction accuracy.

To assess the statistical significance of the mean prediction accuracy per trait in each test dataset, we performed a permutation analysis to generate an empirical null distribution of correlations for each trait and test dataset separately. At each permutation iteration, the trait ratings in a test dataset were shuffled across face images, and the Spearman correlation between the predicted and permuted ratings was computed for each trait. This procedure was repeated 10,000 times to obtain a distribution of the correlations, under the null hypothesis that there is no relationship between facial features and trait ratings. Statistical significance was determined by taking the 95^{th} percentile of the empirical null distribution (p = 0.05). The permutation p-value for each trait was defined as the proportion of the null correlations that were greater than or equal to the observed prediction accuracy. The p-values were corrected for multiple comparisons across the predicted traits using the false discovery rate (FDR) procedure[54].

Variance partitioning analysis. We compared the unique and shared explained variance of each pair of the feature spaces so as not to compromise statistical power if three largely correlated feature spaces were compared at once. Specifically, for each trait and each pair of feature spaces, we fit three models using the training dataset: one fit the trait ratings to a feature space (e.g., 128 *DCNN-Identity* features), the second fit the trait ratings to a second feature space (e.g., 26 *DCNN-Object* features), and the third fit the trait ratings to both feature spaces (e.g., 154 *DCNN-Identity* and *DCNN-Object* features). These three fitted models were used to predict the trait ratings of the faces in the test dataset. The variance explained (r^2) by each model for each trait was computed by squaring the Pearson correlation between the predicted and actual human ratings while keeping its sign[34]. Finally, the unique variance explained by each of the two compared feature spaces (A and B) and the shared variance explained by both feature spaces were computed as follows:

$$r^2_{uA} = r^2_{A\cup B} - r^2_B$$
$$r^2_{uB} = r^2_{A\cup B} - r^2_A$$
$$r^2_{A\cap B} = r^2_A + r^2_B - r^2_{A\cup B}$$

where r^2_A is the total variance explained by the first model using feature space A, r^2_B is the total variance explained by the second model using feature space B, $r^2_{A\cup B}$ is the total variance explained by the third model using features from both spaces, r^2_{uA} is the unique variance explained by feature space A, r^2_{uB} is the unique variance explained by feature space B, and $r^2_{A\cap B}$ is the shared variance explained by feature spaces A and B. We repeated the above analysis procedure with computing the explained variance by squaring the Spearman correlation instead of the Pearson correlation; results corroborated those obtained with the Pearson correlation.

Semi-partial correlation analysis. The semi-partial correlation measures the relationship between two variables X and Y while statistically controlling for (or partialling out) the effect of a third variable Z on Y (note that, in contrast, the partial correlation controls for the effect of Z on both X and Y). In this analysis, the actual trait ratings provided by the human subjects in the test dataset were used as the variable X, the trait ratings predicted by a trait model for the same set of faces were used as the variable Y, and the trait ratings predicted by a second trait model for the same set of faces were used as the variable Z. To partial out the effect of Z from Y, a simple bivariate regression of Y on Z was performed and the residuals were obtained. These residuals quantified the unique variance in Y that were not associated with or predictable from Z. Finally, we computed the Spearman correlation coefficient between X and the residuals.

Data availability. All data are from publicly available datasets which could be accessed via the links provided in the papers cited.

Acknowledgements
Funded in part by NSF grants BCS-1840756 and BCS-1845958, the Simons Foundation Collaboration on the Global Brain (542941), and the Carver Mead New Adventures Fund.

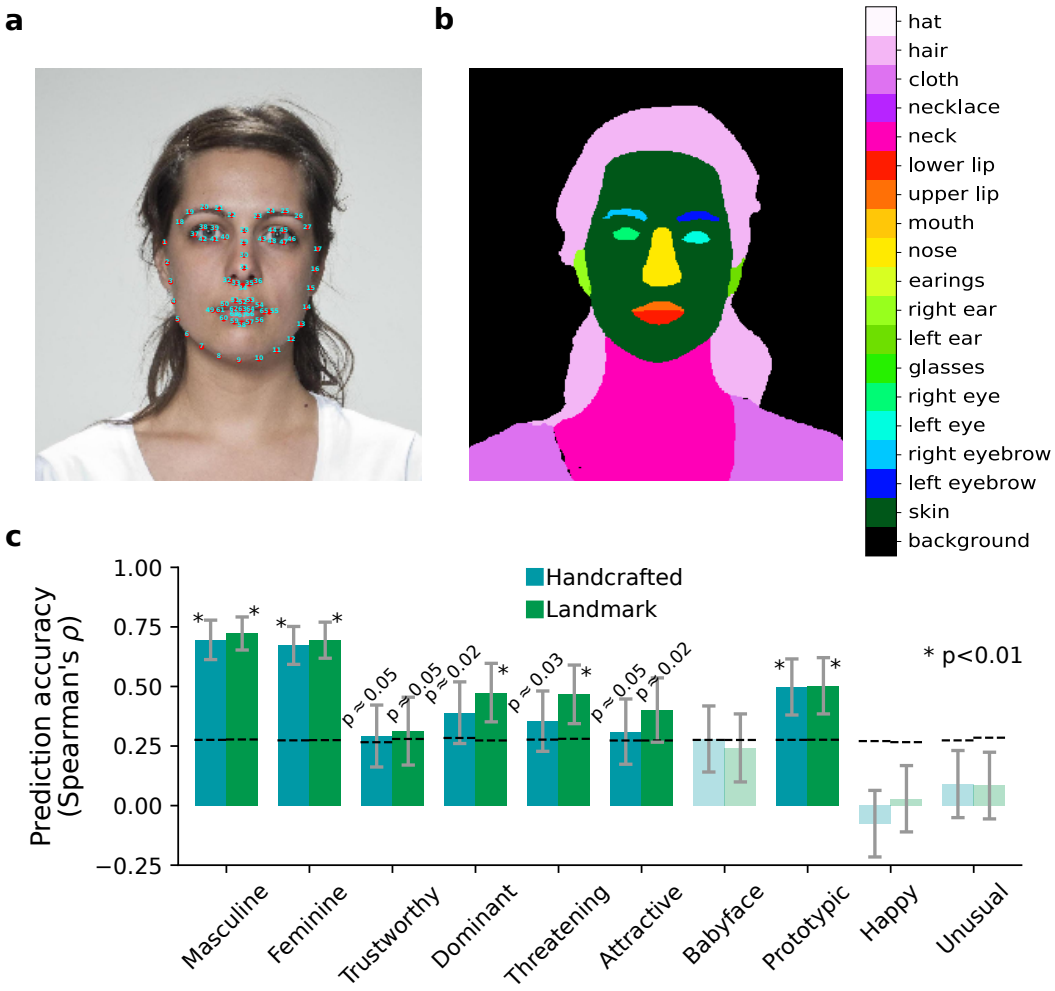
Author contributions
U.K. and R.A. developed the study concept and designed the study; R.A. supervised the experiments and analyses; C.L. performed data collection; U.K. performed data analyses; all authors drafted, revised, and reviewed the manuscript, and approved the final manuscript for submission.

Competing interests
The authors declare no competing interests.

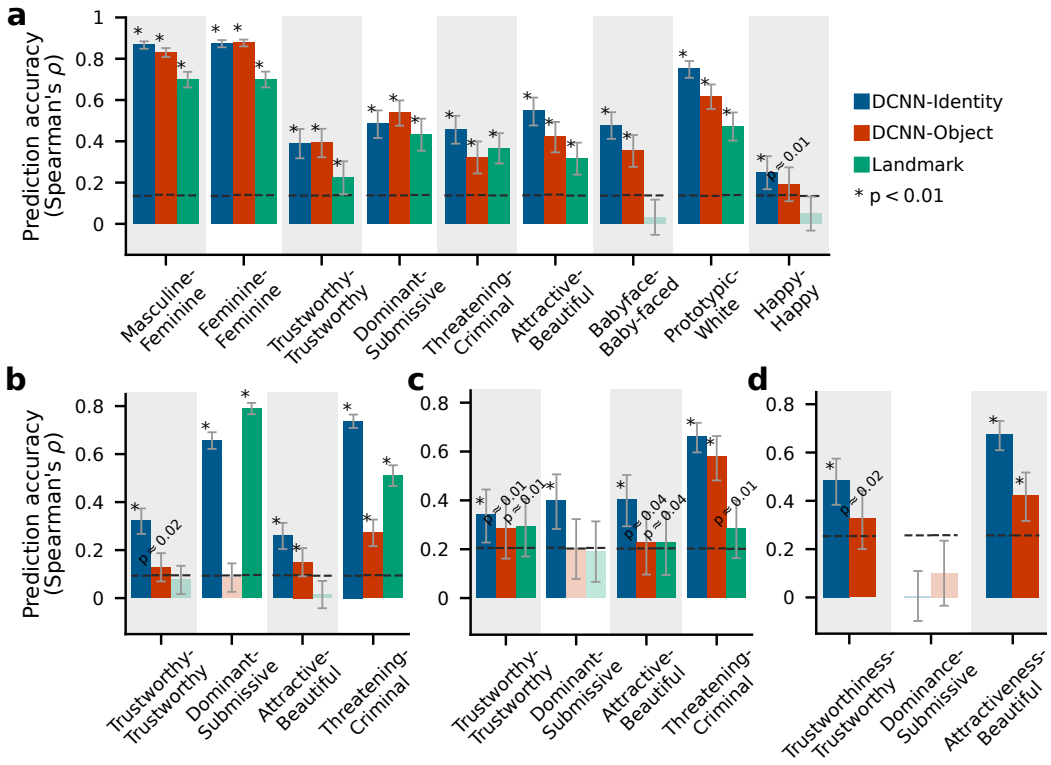
References

1. Sutherland, C. A. M. *et al.* Facial First Impressions Across Culture: Data-Driven Modeling of Chinese and British Perceivers' Unconstrained Facial Impressions. *Pers Soc Psychol Bull* **44**, 521–537 (2018).
2. Willis, J. & Todorov, A. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychol Sci* **17**, 592–598 (2006).
3. Engell, A. D., Haxby, J. V. & Todorov, A. Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in the Human Amygdala. *Journal of Cognitive Neuroscience* **19**, 1508–1519 (2007).
4. Todorov, A. Inferences of Competence from Faces Predict Election Outcomes. *Science* **308**, 1623–1626 (2005).
5. Martin, D. S. Person perception and real-life electoral behaviour. *Australian Journal of Psychology* **30**, 255–262 (1978).
6. Lin, C., Adolphs, R. & Alvarez, R. M. Cultural effects on the association between election outcomes and face-based trait inferences. *PLOS ONE* **12**, e0180837 (2017).
7. Lenz, G. S. & Lawson, C. Looking the Part: Television Leads Less Informed Citizens to Vote Based on Candidates' Appearance. *American Journal of Political Science* **55**, 574–589 (2011).
8. Ahler, D. J., Citrin, J., Dougal, M. C. & Lenz, G. S. Face Value? Experimental Evidence that Candidate Appearance Influences Electoral Choice. *Polit Behav* **39**, 77–102 (2017).
9. Oliviola, C. *et al.* First impressions and consumer mate preferences in online dating and speed-dating. *ACR North American Advances* (2015).
10. Hamermesh, D. S. *Beauty Pays: Why Attractive People Are More Successful* (Princeton University Press, 2011).
11. Gheorghiu, A. I., Callan, M. J. & Skylark, W. J. Facial appearance affects science communication. *PNAS* **114**, 5970–5975 (2017).
12. Wilson, J. P. & Rule, N. O. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. *Psychol Sci* **26**, 1325–1331 (2015).
13. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychol Sci Public Interest* **20**, 1–68 (2019).
14. Lin, C., Keles, U. & Adolphs, R. Four dimensions characterize comprehensive trait judgments of faces. Tech. Rep., PsyArXiv (2019).
15. Lewenberg, Y., Bachrach, Y., Shankar, S. & Criminisi, A. Predicting Personal Traits from Facial Images Using Convolutional Neural Networks Augmented with Facial Landmark Information. *AAAI* **30** (2016).
16. McCurrie, M. *et al.* Convolutional Neural Networks for Subjective Face Attributes. *Image and Vision Computing* **78**, 14–25 (2018).
17. Song, A., Linjie, L., Atalla, C. & Cottrell, G. Learning to see faces like humans: modeling the social dimensions of faces. *Journal of Vision* **17**, 837–837 (2017).
18. Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S. & O'Toole, A. J. Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification. *Cognitive Science* **43**, e12729 (2019).
19. Hill, M. Q. *et al.* Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence* **1**, 522–529 (2019).
20. O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q. & Chellappa, R. Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences* **22**, 794–809 (2018).
21. Bainbridge, W. A., Isola, P. & Oliva, A. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* **142**, 1323–1334 (2013).
22. O'Toole, A. J. *et al.* A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 812–816 (2005).
23. D'Amour, A. *et al.* Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv:2011.03395 [cs, stat]* (2020).
24. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *PNAS* **105**, 11087–11092 (2008).
25. Sutherland, C. A. M. *et al.* Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition* **127**, 105–118 (2013).
26. King, D. E. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* **10**, 1755–1758 (2009).
27. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015).
28. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav Res* **47**, 1122–1135 (2015).
29. Zebrowitz, L. A., Voinescu, L. & Collins, M. A. "Wide-Eyed" and "Crooked-Faced": Determinants of Perceived and Real Honesty Across the Life Span. *Pers Soc Psychol Bull* **22**, 1258–1269 (1996).
30. Oh, D., Dotsch, R., Porter, J. & Todorov, A. Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General* **149**, 323–342 (2020).
31. White, D., Sutherland, C. A. M. & Burton, A. L. Choosing face: The curse of self in profile image selection. *Cognitive Research: Principles and Implications* **2**, 23 (2017).
32. Amos, B., Ludwiczuk, B. & Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. Tech. Rep., CMU-CS-16-118, CMU School of Computer Science (2016).
33. Çukur, T., Huth, A. G., Nishimoto, S. & Gallant, J. L. Functional Subdomains within Scene-Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital Place Area. *J. Neurosci.* **36**, 10257–10273 (2016).
34. Lescroart, M. D. & Gallant, J. L. Human Scene-Selective Areas Represent 3D Configurations of Surfaces. *Neuron* **101**, 178–192.e7 (2019).
35. Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N. & Falvello, V. B. Validation of data-driven computational models of social perception of faces. *Emotion* **13**, 724–738 (2013).
36. Todorov, A., Olivola, C. Y., Dotsch, R. & Mende-Siedlecki, P. Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology* **66**, 519–545 (2015).
37. Walker, M., Schönborn, S., Greifeneder, R. & Vetter, T. The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLOS ONE* **13**, e0193190 (2018).
38. Rule, N. O. *et al.* Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology* **98**, 1–15 (2010).
39. Todorov, A. *Face Value: The Irresistible Influence of First Impressions* (Princeton University Press, 2017).
40. Sutherland, C. A. M. *et al.* Individual differences in trust evaluations are shaped mostly by environments, not genes. *PNAS* **117**, 10218–10224 (2020).
41. Fan, X., Wang, F., Shao, H., Zhang, P. & He, S. The bottom-up and top-down processing of faces in the human occipitotemporal cortex. *eLife* **9**, e48764 (2020).
42. Stolier, R. M. & Freeman, J. B. Functional and Temporal Considerations for Top-Down Influences in Social Perception. *Psychological Inquiry* **27**, 352–357 (2016).
43. Said, C. P., Sebe, N. & Todorov, A. Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* **9**, 260–264 (2009).
44. Knutson, B. Facial expressions of emotion influence interpersonal trait inferences. *J Nonverbal Behav* **20**, 165–182 (1996).
45. Chollet, F. *et al.* Keras. https://keras.io (2015).
46. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (2009).
47. DeBruine, L. & Jones, B. Face research lab london set (2017).
48. Chelnokova, O. *et al.* Rewards of beauty: the opioid system mediates social motivation in humans. *Molecular Psychiatry* **19**, 746–747 (2014).
49. Zebrowitz, L. A. & Collins, M. A. Accurate Social Perception at Zero Acquaintance: The Affordances of a Gibsonian Approach. *Pers Soc Psychol Rev* **1**, 204–223 (1997).
50. Blair, I. V. & Judd, C. M. Afrocentric facial features and stereotyping. In Adams, R. B., Adams Jr, R. B., Ambady, N., Shimojo, S. & Nakayama, K. (eds.) *The science of social vision: The science of social vision*, chap. 18, 306–320 (Oxford University Press, "Oxford", 2011).
51. Lee, C.-H., Liu, Z., Wu, L. & Luo, P. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. 5549–5558 (2020).
52. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55–67 (1970).
53. Holdgraf, C. R. *et al.* Encoding and Decoding Models in Cognitive Electrophysiology. *Front. Syst. Neurosci.* **11** (2017).
54. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

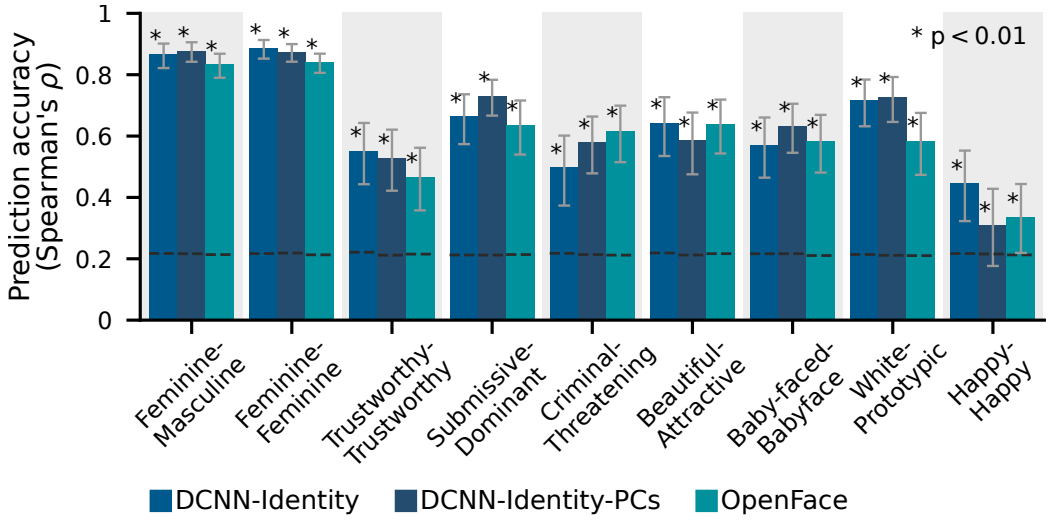
Supplementary Information



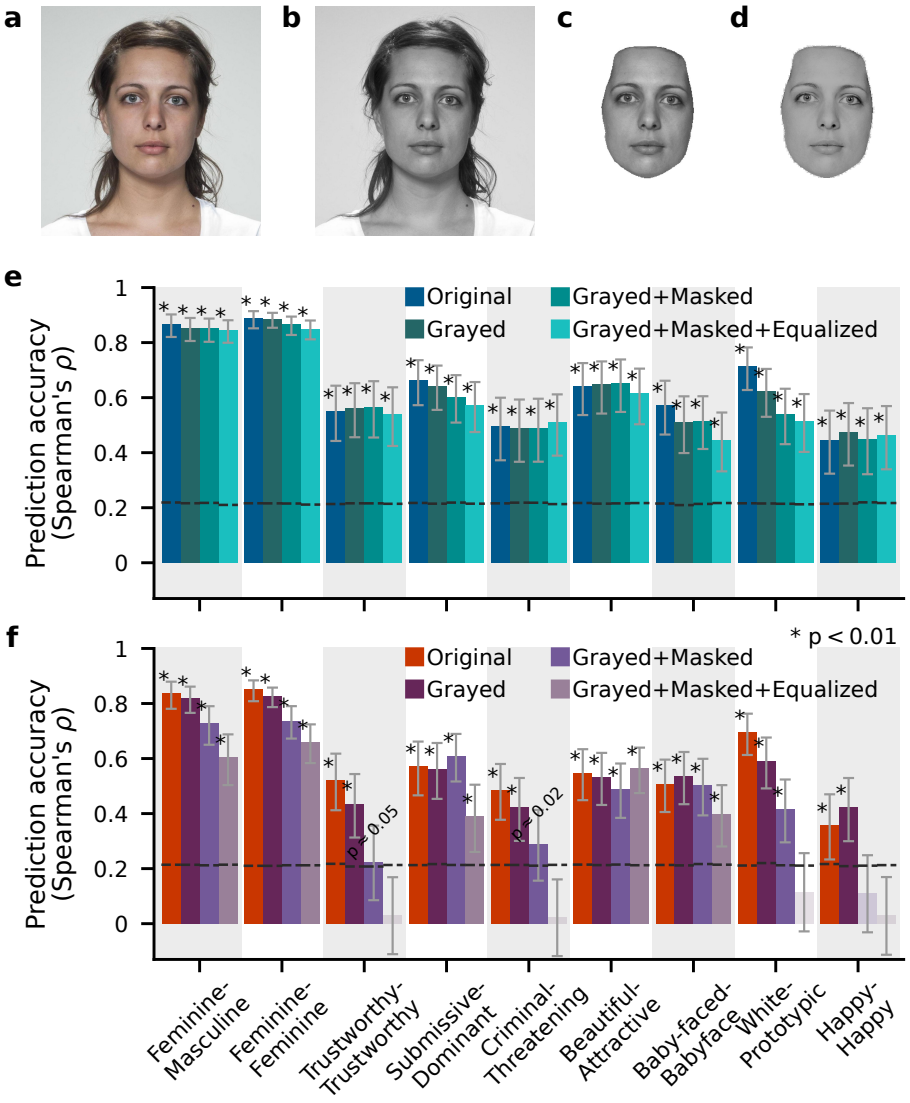
Supplementary Figure 1. Automatically extracted landmark features. a, Automatic detection of facial landmark points. **b,** Automatic detection of facial parts. These landmark points and facial parts were used to automatically measure 30 facial metrics from face images, referred to as Landmark features (e.g., pupillary distance, eye width, median luminance). **c,** Comparison of prediction accuracy from models using automatically extracted Landmark features to models using manually measured facial features provided in the Chicago Face Database[1]. Models were trained and evaluated on the data from the Chicago Face Database using a nested cross-validation procedure (see Methods).



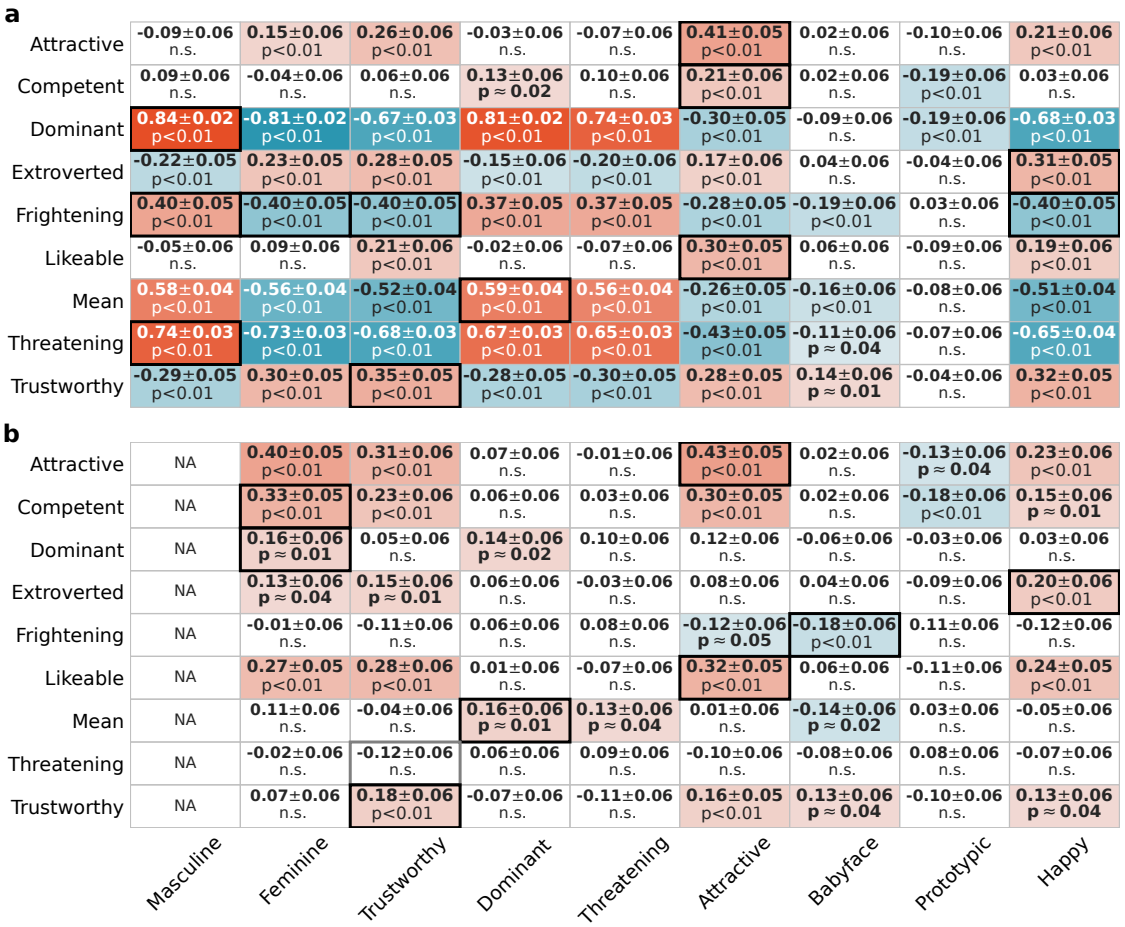
Supplementary Figure 2. Prediction accuracy of models trained on the dataset of Lin et al. (in press). All models were trained on a different dataset than the Chicago Face Database[1], which offered a much larger set of trait ratings in a more diverse sample of subjects[2]. **a,** The prediction accuracy of models tested on the Chicago Face Database (143 nonoverlapping faces between the Chicago Face Database and the current training dataset were used). The bar height indicates the mean prediction and error bars indicate the standard deviations of the mean prediction accuracy across bootstrap samples. Saturated colors, asterisks, and p-values indicate statistically significant predictions ($p < 0.05$, assessed with permutation tests, and FDR corrected); desaturated colors indicate insignificant predictions. Dashed black lines indicate the chance threshold ($p = 0.05$, assessed with permutation test) for the prediction accuracy. **b,** The prediction accuracy of the models tested on the test dataset in Fig. 2b with 300 computer-generated white faces and their trait ratings[3]. **c,** The prediction accuracy of the models tested on the test dataset in Fig. 2c with 66 studio portraits[4]. **d,** The prediction accuracy of the models tested on the test dataset in Fig. 2d with 504 ambient photos of faces in the wild[5].



Supplementary Figure 3. Prediction accuracy and the different aspects of the identity features. Prediction accuracy of DCNN-Identity models in Fig. 2a (light blue) compared to the prediction accuracy of models that used only 30 principal components of the DCNN-Identity features (dark blue)—the same number of regressors as in the Landmark models, and to the prediction accuracy of models using identity features from a different DCNN (turquoise, "OpenFace")[6].



Supplementary Figure 4. Prediction accuracy as a function of low-level image properties. Models were trained and tested on the same training dataset and test dataset as in Fig. 2a. **a,** An example of a face image before any manipulation. **b,** An example of the face image in **(a)** manipulated on colors (i.e., converted to grayscale). **c,** An example of the face image in **(a)** manipulated on hair style (hair was removed). **d,** An example of the face image in **(a)** manipulated on mean luminance (face area luminance histograms were equalized across cropped grayscale face images in the test dataset). **e,** The accuracy of using the model weights obtained from the training dataset (unmanipulated version of the faces) and the DCNN-Identity features extracted from the manipulated versions **(a-d)** of the faces in the test dataset to predict the human subject ratings of the unmanipulated version of the faces in the same test dataset. **f,** Same as **(e)** except that the features were DCNN-Object features.



Supplementary Figure 5. Cross-prediction accuracy for the test dataset in Fig. 2b. a, Cross-prediction accuracy (the Spearman correlations) between the predicted ratings of the faces in the test dataset used in Fig. 2b on nine traits (x-axis) and the human subject ratings of the same set of faces on another nine traits (y-axis)[3]. The saturation of the color indicates the magnitude of the correlation (red for positive, blue for negative). Numbers indicate the mean and standard deviation (across bootstrap samples), and the significance of the correlation (permutation test, FDR corrected) **b,** An example of residual cross-prediction accuracy for traits in the test dataset used in Fig. 2b (y-axis) from eight trait models (x-axis) while controlling for the prediction from the masculine model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). Numbers report the mean bootstrap residual cross-prediction accuracy, bootstrap standard deviation, and significance level computed via permutation tests and FDR corrected. The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes); the highest accuracy per row was highlighted with a solid box (black for significant, grey for insignificant).

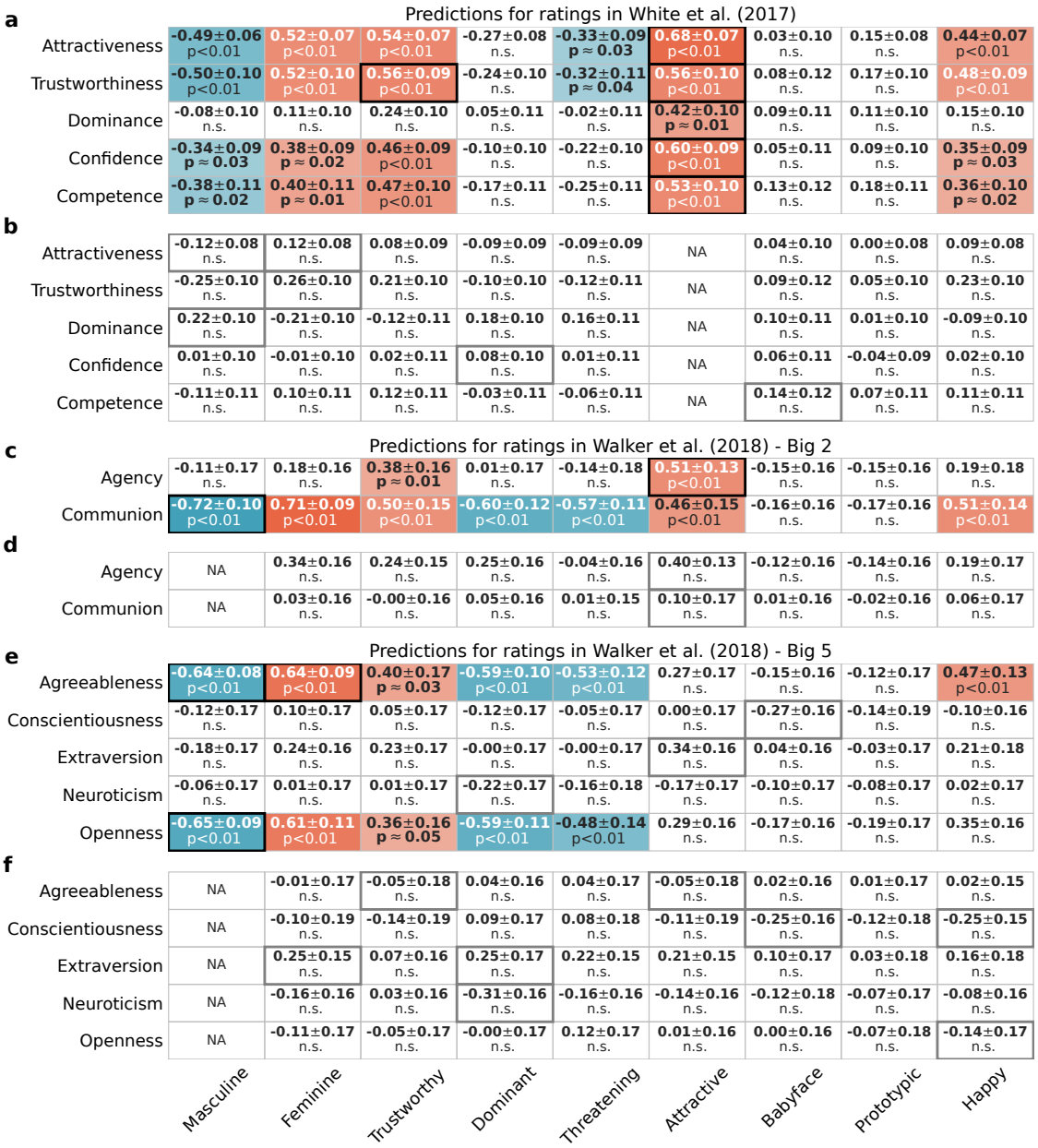
a

	Masculine	Feminine	Trustworthy	Dominant	Threatening	Attractive	Babyface	Prototypic	Happy
Aggressive	0.49±0.10 p<0.01	-0.50±0.10 p<0.01	-0.48±0.09 p<0.01	0.42±0.11 p<0.01	0.33±0.11 p ≈ 0.01	-0.19±0.12 n.s.	0.17±0.12 n.s.	-0.01±0.13 n.s.	-0.42±0.10 p<0.01
Attractive	-0.32±0.12 p ≈ 0.01	0.38±0.12 p<0.01	0.46±0.11 p<0.01	-0.12±0.12 n.s.	-0.14±0.12 n.s.	0.37±0.12 p<0.01	-0.22±0.13 n.s.	0.05±0.12 n.s.	0.38±0.12 p<0.01
Caring	-0.53±0.09 p<0.01	0.55±0.08 p<0.01	0.62±0.07 p<0.01	-0.33±0.11 p ≈ 0.01	-0.37±0.10 p<0.01	0.28±0.12 p ≈ 0.03	-0.22±0.13 n.s.	0.02±0.13 n.s.	0.55±0.09 p<0.01
Confident	-0.16±0.14 n.s.	0.23±0.14 n.s.	0.29±0.13 p ≈ 0.02	0.10±0.13 n.s.	-0.02±0.13 n.s.	0.31±0.13 p ≈ 0.01	-0.10±0.12 n.s.	0.07±0.12 n.s.	0.26±0.14 p ≈ 0.04
Dominant	0.07±0.13 n.s.	0.00±0.14 n.s.	-0.04±0.13 n.s.	0.36±0.12 p<0.01	0.28±0.12 p ≈ 0.03	0.19±0.13 n.s.	-0.11±0.12 n.s.	-0.01±0.13 n.s.	0.04±0.13 n.s.
Emotionally stable	-0.31±0.11 p ≈ 0.02	0.35±0.11 p<0.01	0.43±0.09 p<0.01	-0.08±0.11 n.s.	-0.13±0.11 n.s.	0.20±0.12 n.s.	-0.07±0.13 n.s.	0.11±0.13 n.s.	0.41±0.11 p<0.01
Intelligent	-0.30±0.12 p ≈ 0.02	0.33±0.12 p ≈ 0.01	0.43±0.11 p<0.01	-0.15±0.12 n.s.	-0.14±0.12 n.s.	0.37±0.12 p<0.01	-0.10±0.13 n.s.	0.11±0.13 n.s.	0.38±0.13 p<0.01
Mean	0.27±0.12 p ≈ 0.03	-0.27±0.11 p ≈ 0.03	-0.29±0.10 p ≈ 0.02	0.19±0.12 n.s.	0.14±0.12 n.s.	-0.08±0.12 n.s.	0.07±0.13 n.s.	-0.05±0.13 n.s.	-0.30±0.12 p ≈ 0.02
Responsible	-0.27±0.12 p ≈ 0.03	0.32±0.12 p ≈ 0.01	0.45±0.10 p<0.01	-0.08±0.12 n.s.	-0.12±0.12 n.s.	0.29±0.12 p ≈ 0.02	-0.12±0.13 n.s.	0.17±0.12 n.s.	0.40±0.12 p<0.01
Sociable	-0.51±0.09 p<0.01	0.55±0.09 p<0.01	0.59±0.08 p<0.01	-0.27±0.11 p ≈ 0.03	-0.20±0.12 n.s.	0.36±0.12 p<0.01	-0.19±0.13 n.s.	0.06±0.13 n.s.	0.57±0.09 p<0.01
Threatening	0.61±0.08 p<0.01	-0.61±0.08 p<0.01	-0.66±0.07 p<0.01	0.46±0.09 p<0.01	0.43±0.10 p<0.01	-0.26±0.12 p ≈ 0.04	0.20±0.13 n.s.	-0.09±0.12 n.s.	-0.59±0.08 p<0.01
Trustworthy	-0.35±0.12 p<0.01	0.39±0.11 p<0.01	0.51±0.09 p<0.01	-0.14±0.11 n.s.	-0.19±0.11 n.s.	0.31±0.12 p ≈ 0.02	-0.16±0.13 n.s.	0.19±0.13 n.s.	0.40±0.11 p<0.01
Unhappy	0.15±0.13 n.s.	-0.20±0.13 n.s.	-0.25±0.12 p ≈ 0.05	0.06±0.13 n.s.	0.09±0.13 n.s.	-0.14±0.13 n.s.	-0.04±0.14 n.s.	-0.07±0.13 n.s.	-0.21±0.13 n.s.
Weird	0.24±0.13 p ≈ 0.05	-0.29±0.13 p ≈ 0.02	-0.42±0.11 p<0.01	0.09±0.13 n.s.	0.08±0.13 n.s.	-0.37±0.12 p<0.01	0.16±0.13 n.s.	-0.18±0.12 n.s.	-0.28±0.13 p ≈ 0.03

b

	Masculine	Feminine	Trustworthy	Dominant	Threatening	Attractive	Babyface	Prototypic	Happy
Aggressive	0.27±0.12 n.s.	-0.29±0.12 n.s.	-0.13±0.12 n.s.	0.20±0.12 n.s.	0.09±0.13 n.s.	-0.04±0.13 n.s.	0.16±0.12 n.s.	0.04±0.12 n.s.	NA
Attractive	-0.05±0.13 n.s.	0.11±0.13 n.s.	0.14±0.13 n.s.	0.21±0.13 n.s.	0.14±0.12 n.s.	0.15±0.13 n.s.	-0.21±0.12 n.s.	-0.00±0.12 n.s.	NA
Caring	-0.21±0.12 n.s.	0.24±0.12 n.s.	0.18±0.12 n.s.	0.02±0.12 n.s.	-0.03±0.13 n.s.	0.01±0.12 n.s.	-0.21±0.13 n.s.	-0.05±0.13 n.s.	NA
Confident	0.10±0.12 n.s.	-0.04±0.12 n.s.	0.04±0.13 n.s.	0.36±0.13 n.s.	0.22±0.12 n.s.	0.14±0.13 n.s.	-0.09±0.12 n.s.	0.03±0.12 n.s.	NA
Dominant	0.16±0.12 n.s.	-0.13±0.13 n.s.	-0.16±0.14 n.s.	0.47±0.11 p ≈ 0.01	0.39±0.11 p ≈ 0.04	0.09±0.13 n.s.	-0.11±0.12 n.s.	-0.03±0.13 n.s.	NA
Emotionally stable	-0.00±0.14 n.s.	0.04±0.14 n.s.	0.10±0.12 n.s.	0.23±0.13 n.s.	0.16±0.13 n.s.	-0.02±0.13 n.s.	-0.06±0.13 n.s.	0.05±0.13 n.s.	NA
Intelligent	-0.01±0.12 n.s.	0.05±0.12 n.s.	0.08±0.12 n.s.	0.17±0.13 n.s.	0.14±0.13 n.s.	0.15±0.13 n.s.	-0.09±0.13 n.s.	0.07±0.13 n.s.	NA
Mean	0.16±0.13 n.s.	-0.18±0.13 n.s.	-0.09±0.11 n.s.	0.06±0.12 n.s.	-0.01±0.13 n.s.	-0.02±0.13 n.s.	0.06±0.12 n.s.	-0.03±0.13 n.s.	NA
Responsible	0.01±0.12 n.s.	0.04±0.12 n.s.	0.09±0.12 n.s.	0.27±0.13 n.s.	0.19±0.12 n.s.	0.04±0.13 n.s.	-0.11±0.13 n.s.	0.11±0.12 n.s.	NA
Sociable	-0.16±0.12 n.s.	0.20±0.13 n.s.	0.12±0.12 n.s.	0.11±0.13 n.s.	0.17±0.14 n.s.	0.06±0.13 n.s.	-0.18±0.13 n.s.	-0.01±0.13 n.s.	NA
Threatening	0.32±0.11 n.s.	-0.32±0.12 n.s.	-0.20±0.12 n.s.	0.12±0.12 n.s.	0.11±0.13 n.s.	0.02±0.13 n.s.	0.19±0.12 n.s.	-0.03±0.12 n.s.	NA
Trustworthy	-0.08±0.13 n.s.	0.13±0.12 n.s.	0.19±0.12 n.s.	0.18±0.13 n.s.	0.09±0.13 n.s.	0.11±0.13 n.s.	-0.16±0.13 n.s.	0.15±0.13 n.s.	NA
Unhappy	-0.04±0.14 n.s.	0.04±0.13 n.s.	-0.01±0.12 n.s.	-0.05±0.14 n.s.	-0.09±0.14 n.s.	-0.02±0.12 n.s.	-0.05±0.14 n.s.	-0.03±0.13 n.s.	NA
Weird	0.01±0.13 n.s.	-0.07±0.13 n.s.	-0.19±0.13 n.s.	-0.20±0.14 n.s.	-0.15±0.13 n.s.	-0.19±0.13 n.s.	0.16±0.13 n.s.	-0.15±0.12 n.s.	NA

Supplementary Figure 6. Cross-prediction accuracy for the test dataset in Fig. 2c. a, Cross-prediction accuracy (the Spearman correlations) between the predicted ratings of the faces in the test dataset used in Fig. 2c on nine traits (x-axis) and the human subject ratings of the same set of faces on 14 traits (y-axis)[4]. The saturation of the color indicates the magnitude of the correlation (red for positive, blue for negative). Numbers indicate the mean and standard deviation (across bootstrap samples), and the significance of the correlation (permutation test, FDR corrected). **b,** An example of residual cross-prediction accuracy for traits in the test dataset used in Fig. 2c (y-axis) from eight trait models (x-axis) while controlling for the prediction from the happy model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes).



Supplementary Figure 7. Cross-prediction accuracy across three test datasets. **a,** Cross-prediction accuracy (the Spearman correlations) between the predicted ratings of the faces in the test dataset used in Fig. 2d on nine traits (x-axis) and the human subject ratings of the same set of faces on five traits (y-axis)[5]. **b,** An example of residual cross-prediction accuracy for traits in the test dataset used in Fig. 2d (y-axis) from eight trait models (x-axis) while controlling for the prediction from the attractive model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes). **c,** Cross-prediction accuracy between the predicted ratings of the faces in a fifth out-of-sample test dataset on nine traits (x-axis) and the human subject ratings of the same set of faces on two traits (y-axis)[7]. **d,** An example of residual cross-prediction accuracy for traits in the test dataset as in **(c)** (y-axis) from eight trait models (x-axis) while controlling for the prediction from the masculine model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes). **e,** Cross-prediction accuracy between the predicted ratings of the faces in a sixth out-of-sample test dataset on nine traits (x-axis) and the human subject ratings of the same set of faces on five traits (y-axis)[7]. **f,** An example of residual cross-prediction accuracy for traits in the test dataset as in **(e)** (y-axis) from eight trait models (x-axis) while controlling for the prediction from the masculine model (selected specifically for this test dataset for its largest impacts on cross-predictions across the nine trait models). The significant accuracy was colored (red for positive, blue for negative; more saturated for greater magnitudes).

Supplementary References

1. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav Res* **47**, 1122–1135 (2015).
2. Lin, C., Keles, U. & Adolphs, R. Four dimensions characterize comprehensive trait judgments of faces. Tech. Rep., PsyArXiv (2019).
3. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *PNAS* **105**, 11087–11092 (2008).
4. Oh, D., Dotsch, R., Porter, J. & Todorov, A. Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General* **149**, 323–342 (2020).
5. White, D., Sutherland, C. A. M. & Burton, A. L. Choosing face: The curse of self in profile image selection. *Cognitive Research: Principles and Implications* **2**, 23 (2017).
6. Amos, B., Ludwiczuk, B. & Satyanarayanan, M. Openface: A general-purpose face recognition library with mobile applications. Tech. Rep., CMU-CS-16-118, CMU School of Computer Science (2016).
7. Walker, M., Schönborn, S., Greifeneder, R. & Vetter, T. The Basel Face Database: A validated set of photographs reflecting systematic differences in Big Two and Big Five personality dimensions. *PLOS ONE* **13**, e0193190 (2018).